**PREDICTION OF DISEASE USING MACHINE LEARNING**

Gunda Nikhil Guru Prasad Ramya Sahitya Jakkula  
 *Student ID: 700739893 Student ID: 700742087 Student ID: 700742087  
University of Central Missouri University of Central Missouri University of Central Missouri* [*nxg98930@ucmo.edu*](mailto:nxg98930@ucmo.edu)[*dxg20870@ucmo.edu*](mailto:dxg20870@ucmo.edu)[*rxj74510@ucmo.edu*](mailto:rxj74510@ucmo.edu)

**Abstract:**

Machine learning is the process of programming computers to optimise their performance using examples or prior data. Machine The study of computer systems that learn from data and experience is known as learning. There are two passes in the machine learning algorithm: training and testing. Prediction of a disease based on the symptoms and history of the patient Machine learning technology has been battling for decades. Machine Learning technology provides a good platform in the medical industry for rapidly resolving healthcare challenges. We are using machine learning to keep entire healthcare data. Machine learning technology provides for the rapid analysis of data and the delivery of results. With the use of machine learning technology, clinicians may make better decisions about patient diagnoses and treatment options, leading to an increase in patient outcomes. In this work, we try to embed machine learning capabilities in a single healthcare system. When a diagnosis is not appropriate and disease prediction is implemented using certain predictive machine learning algorithms, healthcare can become smarter. A disease cannot always be recognized in its early stages. This allows the disease prediction to be used successfully. As the saying goes, "prevention is better than cure," disease prediction and outbreak prediction can lead to early disease prevention. The main purpose of this document is to set up a system or as we might say instant medical care that would include symptoms received from multi-sensory devices and other medical data and store them in a health record. This dataset would then be analyzed using K-Mean and C-Means machine learning algorithms to provide maximum-precision results

**Keywords**: K-Mean, Fuzzy C Means, Machine learning, Symptoms, Health Care

**Introduction**

It has been a challenge for decades to use machine learning and data mining techniques to predict disease using patient treatment histories and health data. Numerous studies have used pathology data or medical profiles and data mining techniques to forecast particular diseases. These methods made an effort to forecast the disease's start. Additionally, some methods give predictions about the prevention and development of disease. A trend toward machine learning models that can learn rich, hierarchical representations of raw data with minimum pre-processing and provide more accurate results has been brought about by the recent success of deep learning in a variety of machine learning applications. More focus has been placed on disease prediction from the standpoint of big data analysis with the advancement of big data technology

The basic objective of machine learning in healthcare is to enhance patient care and produce superior results. Making the appropriate diagnosis has become simpler thanks to machine learning. Several effective machine learning algorithms used in predictive analytics help to more effectively identify diseases and treat patients. Health information and treatment history. Later, decisions affecting the patient's health are made using this information that was concealed in the health data. Furthermore, these sectors require improvement through the application of instructive data in healthcare.

One such implementation of machine learning algorithms can be found in the healthcare field. Medical facilities need to be improved so that better decisions can be made about patient diagnoses and treatment options. Machine learning in healthcare helps people process huge and complex medical data sets and then analyzes them into clinical insights. This can then be used by doctors for medical care. Therefore, machine learning, when used in healthcare, can lead to increased patient satisfaction. The kmean algorithm is used to predict diseases based on the patient's treatment history and health data. Symptoms and visual cues are the two main methods to recognise disease. Diagnostic and therapeutic treatments involving healthcare need the use of medical imaging. To do this, fuzzy logic will be employed. with the aid of the clustering technique. Analyzing the project's results with parameters in line with the debate. It's time to make sickness predictions. accuracy for database clustering. The accuracy of fuzzy rules. the level of skill required to develop databases number of patterns in database records. accuracy in drug prescription.

**Motivation**

Medical facilities must be improved in order to make better decisions about patient diagnosis and treatment alternatives. Machine learning in healthcare assists people in processing large and complicated medical information and analysing them to derive therapeutic insights.This can then be used by doctors to provide medical care. As a result, when machine learning is used in healthcare, it can lead to higher patient satisfaction. In this study, we attempt to integrate machine learning functionality in healthcare into a single system. Instead of diagnosis, healthcare may be made smart by implementing disease prediction utilising machine learning predictive algorithms.Some circumstances may arise when an early diagnosis of an illness is not possible. As a result, disease prediction can be implemented efficiently. As widely said “Prevention is better than cure”, prediction of diseases and epidemic outbreak would lead to an early prevention of an occurrence of a disease.

**Objective**

In this research, we used structured and unstructured data in healthcare areas to assess illness risk. The latent factor model approach for reconstructing missing data in medical records collected from the hospital. And, utilising statistical knowledge, we may identify the primary chronic diseases in a specific region and population. We engage hospital specialists to learn about important aspects while dealing with structured data. In the case of unstructured text data, we use the c-mean approach to automatically identify features. For both structured and unstructured data, we offer a K-means and C-Means algorithm.

* to determine a disease
* To help medical students working in pathology labs
* to nurse students
* To doctors for disease-related health issues, and more.

**Related Work**

1. Paper Name: Disease phenol type similarity improves the prediction of novel disease associated micro RNAs Author: Duc-Hau Le[1]

Several studies have demonstrated the importance of miRNAs (microRNAs) in human disease, and a variety of computational approaches have been developed to anticipate such connections by ranking candidate microRNAs according to their relevance to a disease. Network-based techniques are gaining traction since they make efficient use of the "disease module" idea in functionally similar miRNA networks. RWRMDA is a state-of-the-art algorithm-based Random Walk with Restart (RWR) technique on a functional similarity network miRNA. Because the concept of "disease module" also exists in protein interaction networks, the application of this method was motivated by its performance in predicting disease genes. Several more algorithms have also been designed for the prediction of disease genes. Regardless, they have not been used yet for disease microRNA prediction

1. Paper Name: Defining Disease Phenotypes in Primary Care Electronic Health Records by a Machine Learning Approach: A Case Study in Identifying Rheumatoid Arthritis Author: Shang-Ming Zhou[2]

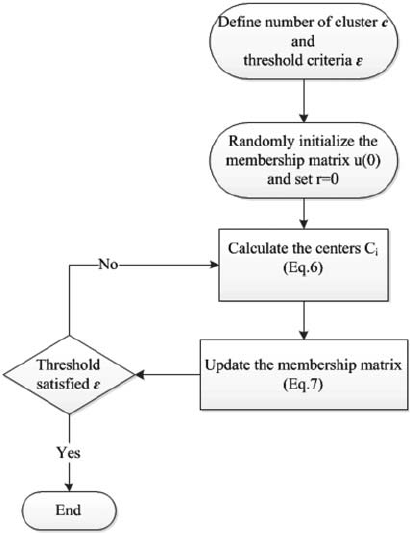
Variables were chosen by comparing the relative frequencies of Read codes in the primary care dataset associated with disease cases versus non-disease controls (disease/non-disease based on secondary care diagnosis); ii) predictors/associated variables were reduced using a Random Forest method; and iii) decision rules were inferred from decision tree models. The suggested method was then thoroughly verified on a separate dataset and its performance was compared to two existing deterministic algorithms for RA created with expert clinical knowledge.

1. Paper Name: Design and Implementing Heart Disease Prediction Using Naives Bayesian Author: Anjan Nikhil Repaka, Sai Deepak Ravikanti

The suggested method includes the following phases: dataset selection, application-based user registration and login, Bayesian Navy classification, prediction, and stable data transfer utilising AES (Advanced Encryption Standard). The results are then generated. The work elaborates and offers many information abstraction strategies by utilising data mining tools used for heart disease prediction. The findings suggest that the developed diagnostic framework accurately predicts risk factors for heart disease.[3]

**Proposed Framework**

**Fuzzy C Means flow**

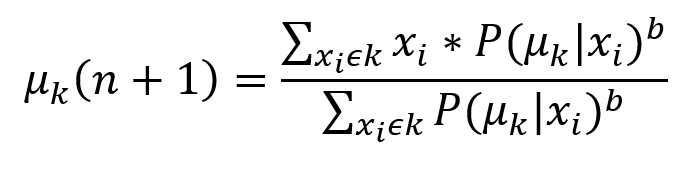


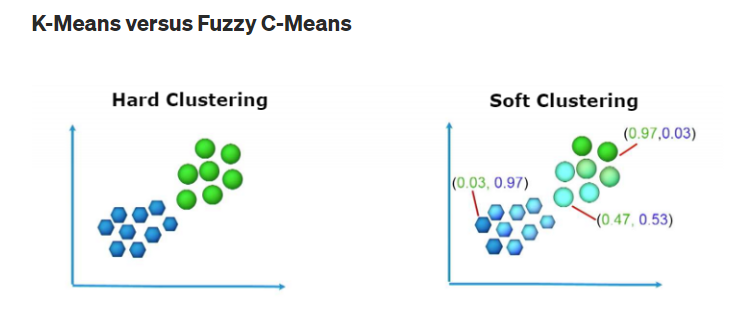
**Fuzzy c-means algorithm**

Attribution to a cluster: In fuzzy clustering, each point has a probability of belonging to each cluster, rather than belonging to only one cluster as in standard k-means. Each point in Fuzzy-C Means clustering has a weighting connected with a certain cluster, therefore a point does not sit "in a cluster" as much as it has a weak or strong relationship to the cluster, which is determined by the inverse distance to the centre of the cluster.

Speed: Fuzzy-C means tend to run slower than K means since they do more work. Each point is evaluated with each cluster, and each evaluation involves more operations. K-Means only needs to calculate distance, whereas fuzzy c means must also calculate distance. must perform full inverse-distance weighting

Personal Opinion: When it comes to extended clusters, FCM/Soft-K-Means is "less foolish" than Hard-K-Means (when points otherwise consistent in other dimensions tend to scatter along a particular dimension or two).

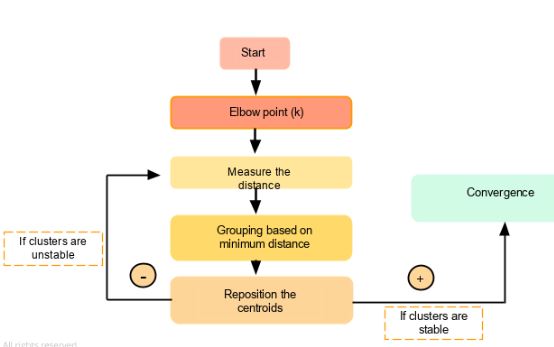




**Flow**

* Assume that there are k clusters.
* Initialization: Randomly initialise the k-means k associated with the clusters and compute P(point xi has label k|xi, k) the chance that each data point xi is a member of a specific cluster k.
* Iteration: Recalculate the cluster centroid as the weighted centroid given the membership probabilities of all data points xi:
* Ending: Iterate until convergence or a user-specified number of iterations is reached (the iteration may be trapped at some local maxima or minima).

**K-Mean flow**



**Methodology**

Machine Learning is the domain that predicts using prior data. Machine Learning is a computer system concept in which the Machine Learning model learns from data and experience. The machine learning algorithm is divided into two stages: 1) Training and 2) Testing Machine learning technology has been striving for decades to forecast disease based on a patient's symptoms and history. Machine Learning Technology can be used to efficiently handle healthcare challenges.

We are using entire machine learning methods to monitor the health of our patients. We can use ML models to develop models that quickly clean and process data and offer results. Using this approach, clinicians will be able to make sound judgements about patient diagnoses and, as a result, provide appropriate treatment will be delivered to the patient, which improves patient healthcare services Healthcare is an excellent illustration of how machine learning may be applied in the medical industry.

**System configuration**

This project may be run on standard hardware. We ran the entire project on an Intel I5 processor with 8 GB RAM and a 2 GB Nvidia Graphic Processor. It also has two cores that run at 1.7 GHz and 2.1 GHz. The first half of the process is the training phase, which takes about 10-15 minutes, and the second part is the testing phase, which just takes a few seconds to generate predictions and calculate accuracy.

**Hardware Requirements:**

• RAM: 4 GB

• Storage: 500 GB

• CPU: 2 GHz or faster

• Architecture: 32-bit or 64-bit

**Software requirements**

• Python 3.5 in Google Colab is used for data pre-processing, model training and prediction.

• Operating System: windows 7 and above or Linux based OS or MAC OS.

**Data Description**

A dataset to provide the students a source to create a healthcare related system.  
Get\_dummies processed file will be available at <https://www.kaggle.com/rabisingh/symptom-checker?select=Training.csv>

There are columns containing diseases, their symptoms , precautions to be taken, and their weights. This dataset can be easily cleaned by using file handling in any language. The user only needs to understand how rows and columns are arranged.

**Result Analysis**

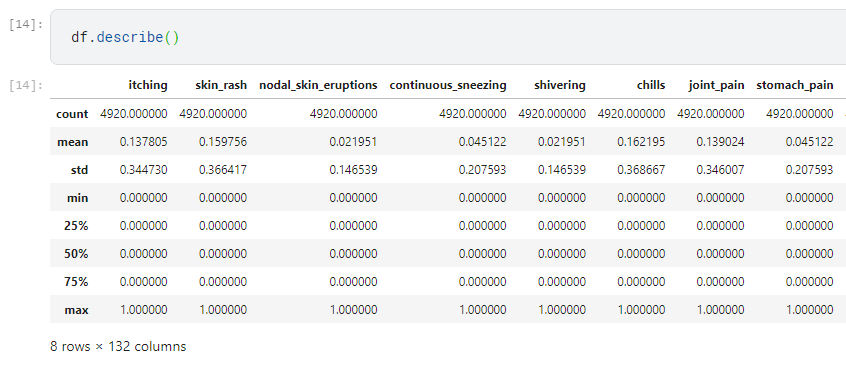
Data visualisation is the discipline of attempting to understand data by displaying it in a visual context in order to highlight patterns, trends, and connections that might otherwise go undetected.Python has a number of excellent graphing packages that are jam-packed with useful functionality. Python provides a great library for creating dynamic or highly customizable charts.

To get a little overview, here are a few popular plotting libraries:

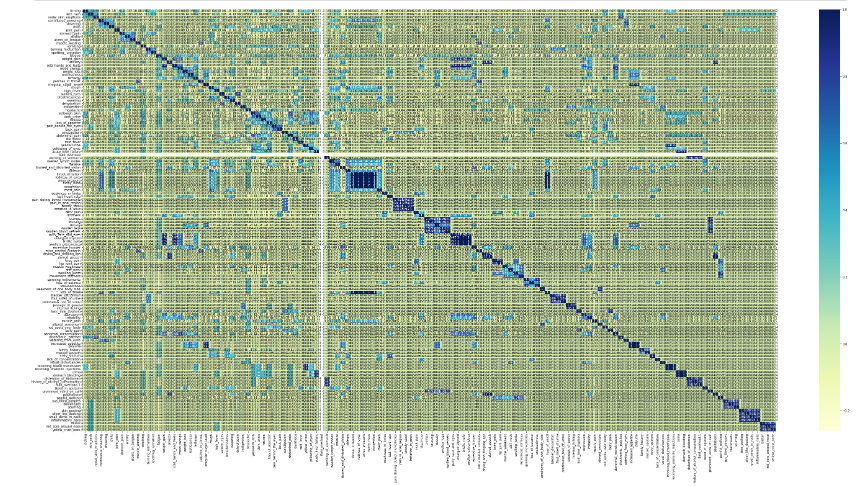
* [**Matplotlib:**](https://matplotlib.org/)low level, provides lots of freedom
* [**Pandas Visualization:**](https://pandas.pydata.org/pandas-docs/stable/visualization.html)easy to use interface, built on Matplotlib
* [**Seaborn:**](https://seaborn.pydata.org/)high-level interface, great default styles
* [**plotnine:**](https://plotnine.readthedocs.io/en/stable/)based on R’s ggplot2, uses Grammar of Graphics
* [**Plotly:**](https://plot.ly/python/)can create interactive plots

**Preliminary Results**

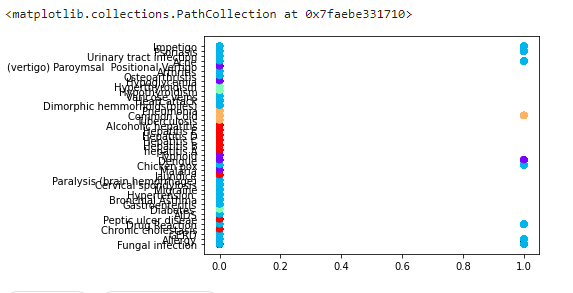
Data Stats



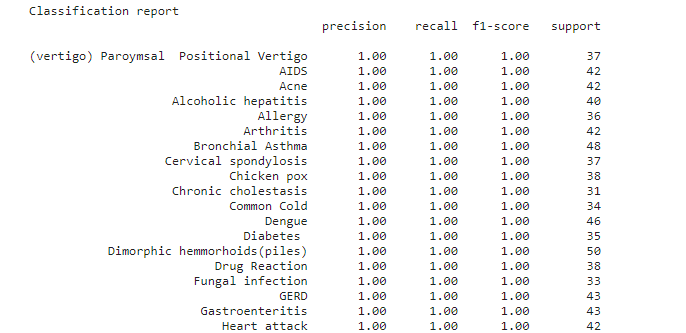
**Correlation graph**



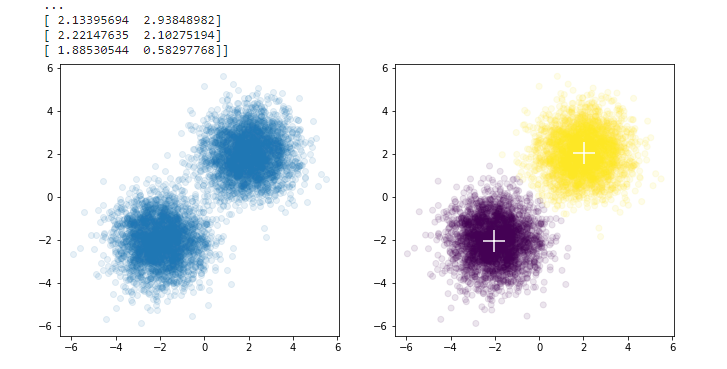
K-Means



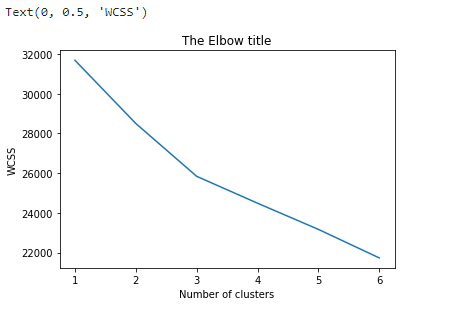
**Classification Report**



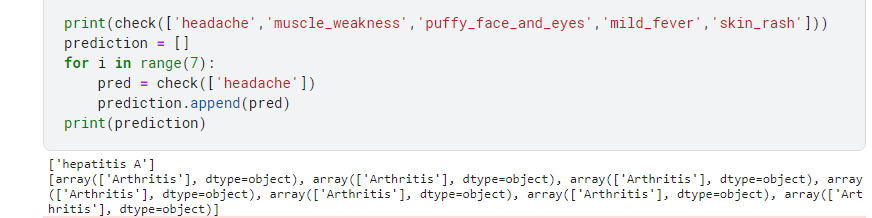
**CMeans-Clusters**



**K-Means Cluster**



**Prediction using symptoms**



**References/Bibliography**

[1] "2015 2nd National Foundation for Science and Technology Development Conference on Information and Computer Science" Duc-Hau Le, "Disease phenotypic similarity improves the prediction of novel diseaseassociated microRNAs"

[2] Shang-Ming Zhou, "Defining Disease Phenotypes in Primary Care Electronic Health Records by a Machine Learning Approach: A Case Study in Identifying Rheumatoid Arthritis", doi:10.1371/journal.pone.0154515"

[3] Anjan Nikhil Repaka, "Design and Implementing Heart Disease Prediction Using Naives Bayesian," in Proceedings of the Third International Conference on Trends in Electronics and Informatics (ICOEI 2019); IEEE Xplore Part Number: CFP19J32-ART; ISBN: 978-1-5386-9439-8

[4] Min Chen, Yixue Hao, Kai Hwang, Fellow, IEEE, Lu Wang, and Lin Wang "Disease Prediction by Machine Learning from Healthcare Communities" (2017).

Mr. Chala Beyene and Prof. Pooja Kamat, "Survey on Prediction and Analysis of Heart Disease Occurrence Using Data Mining Techniques," International Journal of Pure and Applied Mathematics, 2018.

[5] P. Groves, B. Kayyali, D. Knott, and S. V. Kuiken, 2016. "The 'big data' revolution in healthcare: Accelerating value and innovation."

[6] S. Patel and H. Patel, "Survey of data mining techniques utilised in the healthcare domain," International Journal of Informatics Science and Technology, Vol. 6, pp. 53-60, March 2016.